РАСПОЗНАВАНИЕ ЗВУКОВ РЕЧИ ЧЕЛОВЕКА НА ОСНОВЕ МЕТОДА АППРОКСИМАЦИИ

В. В. Митянок, Н. В. Коновалова (Пинск, Беларусь)
mitsianok@tut.by

Для автоматического распознавания речи человека разработан ряд компьютерных программ, наиболее известными из которых являются программы серии Dragon. Однако в настоящее время прогресс в данном направлении явно замедлился. Пользователи отмечают ряд недостатков программ Dragon, из которых наиболее существенны следующие:

1. Необходимость предварительного обучения программы голосу пользователя.

2. Влияние акцента пользователя на надежность распознавания.

3. Зависимость надежности распознавания даже от слабых посторонних шумов.

4. Даже идеально чистый голос при полном отсутствии посторонних шумов не дает 100-процентную надежность распознавания.

Скорее всего, замедление прогресса связано с тем, что идеи, лежащие в основе нынешних систем распознавания, уже выработаны, и для дальнейшего продвижения вперед требуется привлечь новые.

Для распознавания звуков речи человека представляется естественным использовать преобразования Фурье. Это связано с тем, что кривые звукового давления большинства звуков представляют собой периодические (или почти периодические) функции времени. Однако преобразования Фурье обладают рядом недостатков, из которых отметим следующие, важные для проблемы распознавания речи человека.

1. Размытость получаемого спектра даже для идеального синусоидального сигнала, но взятого на ограниченном отрезке времени. Степень размытости зависит от длительности сигнала. (Это обстоятельство является математической подоплекой хорошо известного в квантовой механике соотношения неопределенностей).

2. Наличие фальшивых максимумов в спектре сигналов.

3. Неустойчивость спектра при наличии помех и (или) неизбежных дрожаний голоса.

4. Вытекающая из предыдущего пункта трудность расшифровки спектра.

5. Плохая идентификация малоинтенсивных мод.

В связи с вышеизложенными проблемами, в [1–3], в противовес методу преобразований Фурье, был предложен метод аппроксимации для разложения отдельных звуков речи человека на отдельные моды, параметры которых медленно (по сравнению с несущей частотой) меняются со временем. Метод основан на функционале

$$S = \sum_{i=1}^{n}(y_i - b_{0i} - \sum_{k=1}^{l_1} a_{ki}\sin(\omega_k i) - \sum_{k=1}^{l_1} b_{ki}\cos(\omega_k i))^2 + \\ +\alpha\sum_{i=1}^{n-1}(b_{0,i} - b_{0,i+1})^2 + \alpha\sum_{k=1}^{l_1}\sum_{i=1}^{n-1}(b_{k,i} - b_{k,i+1})^2 + \alpha\sum_{k=1}^{l_1}\sum_{i=1}^{n-1}(a_{k,i} - a_{k,i+1})^2. \quad (1)$$

где y_i – наблюдаемая величина звукового давления в точке номер i (значение аппроксимируемой функции), l_1 – количество мод, ω_k – их несущие частоты, a_{ki}, b_{ki}, b_{0i} – соответственно их дрейфующие синус-, косинус- – амплитуды и начало отсчета, n – количество оцифрованных точек на отобранном для изучения отрезке звуковой кривой, α – параметр метода. У дрейфующих амплитуд и начала отсчета первый индекс соответствует номеру моды, второй – порядковому номеру точки звуковой кривой.

Если бы в (1) других слагаемых, кроме первого, не было, то тогда аппроксимация про-водилась бы в каждой точке звуковой кривой по отдельности, она была бы идеальной (т. е. остаточная невязка равнялась бы нулю). Но при этом найденные амплитуды испытывали бы резкие скачки от точки к точке, что не соответствует действительности. Поэтому в (1) введены дополнительные слагаемые, ограничивающие эти скачки. Эти слагаемые сопровождаются положительным множителем α, ответственным за относительную важность слагаемых.

Для упрощения получающихся впоследствии выражений определим новые величины

$$\psi_{j,i} = \sin(\omega_j i), \psi_{j+l_1,i} = \cos(\omega_j i), \psi_{l,i} = 1,\ j = 1,\ldots,l_1,\ i = 1,\ldots,n. \quad (2)$$

Оформим все дрейфующие амплитуды и нуль отсчета общим списком, записав их в матрицу – столбец X: l_1 раз, в порядке возрастания номеров несущих частот, амплитуды синус- волн, каждая по n элементов, соответствующих n оцифрованным моментам времени, всего

$n1$ чисел. Затем, сразу же за ними – амплитуды косинус- волн, в том же порядке и количестве. После них – дрейфующее начало отсчета, в количестве n чисел:

$$X[i+(k-1)n] = a_{k,i},\ X[nl_1+i+(k-1)n] = b_{k,i}, i = 1,\dots,n,\ k = 1,\dots,l_1. \tag{3}$$

$$X[i+2nl_1] = b_{0,i}, \quad i = 1,\dots,n, \tag{4}$$

Теперь функционал (1) может быть записан в виде

$$S = \sum_{i=1}^{n}(y_i - \sum_{k=1}^{l}\psi_{k,i}X_{i+(k-1)n})^2 + \alpha\sum_{k=1}^{l}\sum_{i=1}^{n-1}(X_{i+(k-1)n}, -X_{i+1+(k-1)n})^2, \tag{5}$$

где $l = 2l_1 + 1$. Согласно методу наименьших квадратов, функционал (5) следует продифференцировать по всем X и затем приравнять полученные частные производные нулю. В результате получим систему уравнений для нахождения тех значений X, которые обеспечивают минимальное значение невязки S. Эта система имеет вид

$$\psi_{k,1}\sum_{m=1}^{l}\psi_{m,1}X_{1+(m-1)n} + \alpha(X_{1+(k-1)n} - X_{2+(k-1)n}) = \psi_{k,1}y_1,\ k = 1,\dots,l, \tag{6}$$

$$\psi_{k,n}\sum_{m=1}^{l}\psi_{m,n}X_{mn} + \alpha(X_{kn} - X_{kn-1}) = \psi_{k,n}y_n,\ k = 1,\dots,l, \tag{7}$$

$$\psi_{k,i}\sum_{m=1}^{l}\psi_{m,i}X_{i+(m-1)n} + \alpha(2X_{i+(k-1)n} - X_{i+1+(k-1)n} - X_{i-1+(k-1)n}) = \psi_{k,i}y_i, \tag{8}$$

где $k = 1,\dots,l, i = 1,\dots,n-1$. Упорядочивая в (6)–(8) все неизвестные X по возрастанию значений их индексов, запишем полученную систему линейных алгебраических уравнений в матричном виде

$$\sum_{j=1}^{nl}A_{i,j}X_j = B_i, \quad i = 1,\dots,n_1. \tag{9}$$

Из (6)–(9) следует, что симметричная матрица может быть представлена в блочном виде

$$A = \begin{pmatrix} Z_{11} & Z_{12} & . & Z_{1l} \\ Z_{21} & Z_{22} & . & Z_{2l} \\ . & . & . & . \\ Z_{l1} & Z_{l2} & . & Z_{ll} \end{pmatrix} \tag{10}$$

Каждый из блоков $Z_{i,j}$, $i,j = 1,\ldots,l$, имеет размер $n \times n$. Неравные нулю элементы блоков Z, образующих матрицу A, есть

$$(Z_{mm})_{i,i} = (\psi_{m,i})^2 + 2\alpha, \quad m = 1,\ldots,l, \quad i = 2,\ldots,n-1, \tag{11}$$

$$(Z_{mm})_{1,1} = (\psi_{m,1})^2 + \alpha, \quad (Z_{mm})_{n,n} = (\psi_{m,n})^2 + \alpha, \quad m = 1,\ldots,l, \tag{12}$$

$$(Z_{mm})_{i,i-1} = -\alpha, \quad m = 1,\ldots,l, \quad i = 1,\ldots,n-1, \tag{13}$$

$$(Z_{mm})_{i-1,i} = -\alpha, \quad m = 1,\ldots,l, \quad i = 2,\ldots,n, \tag{14}$$

$$(Z_{m,k})_{i,i} = \psi_{k,i}\psi_{m,i}, \quad k,m = 1,\ldots,l, \quad k \neq m, \quad i = 1,\ldots,n. \tag{15}$$

Матрица-столбец в (9) также может быть представлена в блочном виде – l столбцов по n элементов в каждом.

$$B_{i+n(k-1)} = y_i\psi_{k,i}, \quad k = 1,\ldots,l, \quad i = 1,\ldots,n.(16) \tag{16}$$

Если аппроксимируемой функцией является сумма некоторого числа синусоид с произвольными частотами, амплитудами и фазами, и если их несущие частоты известны, то решение (9), перелицованное затем перестановками (3)–(4), получается в виде независящих от времени (то есть от второго индекса) чисел a_{ki}, b_{ki}, b_{0i}. Причем независимо ни от длительности сигнала, ни от параметра метода α. Это – достоинство метода. Недостатком метода является то, что при большом n система уравнений получается непомерно большой, что, в свою очередь, приводит к значительным затратам компьютерного времени. Ниже предложим способы ослабления этого недостатка.

Если найти матрицу A^{-1}, обратную к матрице A, то решение (9) с учетом (16) может быть записано в виде

$$\begin{aligned} X_j = \sum_{k=1}^{nl} A^{-1}_{j,k}B_k = \sum_{i=1}^{n}\sum_{m=1}^{l} A^{-1}_{j,i+n(m-1)}B_{i+n(m-1)} = \\ = \sum_{i=1}^{n}\sum_{m=1}^{l} A^{-1}_{j,i+n(m-1)}y_i\psi_{m,i} = \sum_{i=1}^{n} y_i \sum_{m=1}^{l} A^{-1}_{j,i+n(m-1)}\psi_{m,i}. \end{aligned} \tag{17}$$

Выражение $\sum_{m=1}^{l} A^{-1}_{j,i+n(m-1)}\psi_{m,i}$, входящее в (17), можно подсчитать заранее, что намного сократит компьютерное время, необходимое для расчета дрейфующих амплитуд и начала отсчета. Однако, если n и l_1 велики, то необходимое для расчетов компьютерное время может, несмотря на вышеприведенный ход, оказаться все еще слишком большим. В этом случае возникает идея разрезать изучаемый отрезок звуковой кривой на фрагменты, каждый из них обработать по вышеописанной методике, а затем состыковать между собой моды с одинаковыми номерами, но принадлежащие разным фрагментам. При этом

нужно учитывать сдвиг начала отсчета времени в разных фрагментах.

Рассмотрим подробнее вопрос о фрагментах со сдвинутым началом отсчета. Пусть некоторая функция (фрагмент) $y_i = y(i)$ задана уравнением

$$y_i = b_0 + \sum_{i=1}^{l_1} a_k \sin(\omega_k i) + \sum_{i=1}^{l_1} b_k \cos(\omega_k i), \tag{18}$$

где i изменяется от $N+1$ до $N+n$, а N имеет смысл сдвига начала отсчета i. В (18) заменим i на $i+N$, а затем $y_i + N$ на $y_i^{'}$. Получим

$$y_i^{'} = y_{i+N} = b_0 + \sum_{i=1}^{l_1} a_k \sin(\omega_k(i+N)) + \sum_{i=1}^{l_1} b_k \cos(\omega_k(i+N)),\ i = 1, \ldots, n. \tag{19}$$

С другой стороны, $y_i^{'}$ может быть записано как

$$y_i^{'} = b_0 + \sum_{i=1}^{l_1} a_i^{'} \sin(\omega_k i) + \sum_{i=1}^{l_1} b_i^{'} \cos(\omega_k i), \quad i = 1, \ldots, n. \tag{20}$$

где a_i^i и $b_i^{'}$ относятся уже к сдвинутому фрагменту. Раскрывая выражение (19) и сопоставляя с (20) легко увидеть, что амплитуды волн для одного и того же участка, но рассматриваемого с двух разных точек зрения связаны между собой соотношениями

$$a_k = a_k^{'} \cos(\omega_k N) + b_k^{'} \sin(\omega_k N), \tag{21}$$

$$b_k = -a_k^{'} \sin(\omega_k N) + b_k^{'} \cos(\omega_k N). \tag{22}$$

Таким образом, если взять не весь участок звуковой кривой, а лишь его фрагмент, начинающийся при $i = N+1$, то для стыковки его с предыдущими фрагментами следует для найденных амплитуд совершить преобразования (21)–(22).

Однако практика применения вышеописанного метода показала, что прямая стыковка фрагментов реальной звуковой кривой дает заметный скачок в точке их соединения. Вызвано это краевыми эффектами. Чтобы их нивелировать, фрагменты звуковой кривой следует нарезать с некоторым перекрытием. Ниже приведен пример расчета динамики амплитуды одной из мод звука "A", проведенный с частичным перекрытием по времени.

После изучения более 5000 образцов фрагментов звуковых кривых было найдено, что удовлетворительная их стыковка достигается, если

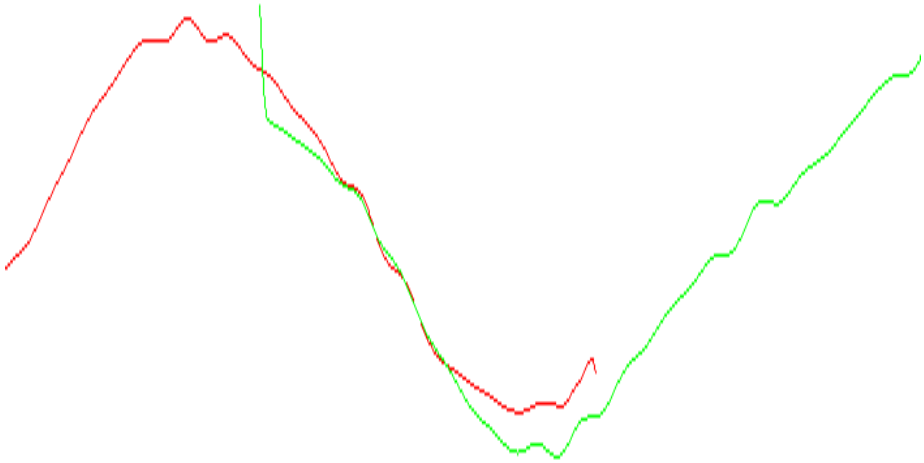

Рисунок 1 – Результаты расчета динамики амплитуды одной из мод звука "A" по двум соседним фрагментам с частичным перекрытием. Длина каждого из фрагментов – 1000 точек при частоте дискретизации 44100 Гц, длина участка перекрытия – 500 точек. По горизонтальной оси отложено время, по вертикальной – величина звукового давления.

длина каждого из фрагментов – не меньше, чем 1000 точек, а зона перекрытия – 500 точек. После расчетов, для стыковки, следует применить преобразования (21)–(22), а затем обрезать с каждого из краев фрагментов по половине длины участка перекрытия, то есть по 250 точек. Таким путем можно анализировать звуковые кривые, имеющие длительность 100–200 и более тысяч точек, т. е., при частоте дискретизации 44100 Гц, более 2–5 секунд.

Список литературы

[1] Митянок В.В. *Метод аппроксимации для нахождения числовых характеристик некоторых низкочастотных звуков человеческой речи* //Электронный журнал «Техническая акустика», http://www.ejta.org 2008, 15 СП-6.

[2] Митянок В.В. *Определение числовых характеристик высокочастотных звуков речи на основе аппроксимации гармоническими функциями* //Известия НАН Беларуси, 2009. №2. С. 111–118.

[3] Митянок В.В. *Применение аппроксимации для автоматического распознавания речи* // Тезисы 6-й межд. науч. конф. «Информационные системы и технологии» IST-2010. Мн.: Инфопарк. 25-25 ноября 2010. С. 201–204.